A Security-Aware Collaborative Decision Optimization Algorithm for Multi-UAV Systems

LI Yizhe[1], XIE Chenyu[1], LIU Shuming[2], WAN Ziheng[1], WEI Xintan[2], DONG Lu[1*]

(1. School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China; 2. School of Computer Science and Engineering, Southeast University, Naning 211189, China)

Abstract: This paper addresses the dual challenge of security and robustness in collaborative decision-making for multi-UAV systems operating in dynamic and adversarial environments, where traditional approaches that decouple safety mechanisms from control policies often fail under anomalies. To this end, we propose adaptive security control with adversarial-resilient endogenous strategy (ASC-ARES), a novel framework grounded in "security by design" and "security left shift" principles that systematically embeds multi-layer constraints, including biconnected topology control, physical collision avoidance, and energy management, into deep reinforcement learning via structured state modeling and reward shaping. Methodologically, ASC-ARES extends the deep deterministic policy gradient (DDPG) algorithm to handle hybrid action spaces through a dual-head policy network for joint optimization of three-dimensional continuous attitude and discrete yaw actions. It further integrates a centroid-guided biconnectivity control algorithm to enable proactive network connectivity awareness and constructs a mean opinion score (MOS)-driven multi-objective adaptive reward mechanism to synergistically optimize quality of experience (QoE), network resilience, safety, and energy efficiency. Experimental results demonstrate that ASC-ARES achieves superior convergence and stability, maintaining an MOS fluctuation rate of only 0.36% and a biconnectivity success rate of 99.98%. Under fast gradient sign method (FGSM), projected gradient descent (PGD), and strong noise interference ($\epsilon=2.0$), the system exhibits exceptional topology reconstruction and state recovery capabilities, with an average performance restoration rate exceeding 80% after interference removal. Ablation studies confirm that the topology control module improves service quality by 59%, while the repulsion mechanism reduces collision risk by 85%. These findings establish ASC-ARES as an effective paradigm for achieving integrated performance-security co-optimization in resource-constrained multi-agent systems.

Highlights:

1. ASC-ARES: A security-by-design framework that endogenously embeds topology, collision, and energy constraints into deep reinforcement learning for multi-UAV coordination.
2. Integration of centroid-guided biconnected topology control with hybrid-action deep deterministic policy gradient (DDPG) enables joint optimization of connectivity, safety, and energy efficiency.
3. Achieves 99.98% network biconnectivity and only 0.36% mean opinion score (MOS) fluctuation under dynamic operational conditions.
4. Demonstrates over 80% recovery in system performance after fast gradient sign method (FGSM) and projected gradient descent (PGD) attacks, with ablation studies confirming an 85% reduction in collision risk.

Key words: multi-UAV systems; deep reinforcement learning; collaborative decision-making; security by design; topology control

Foundation item: National Natural Science Foundation of China (No.62576100).

Received: 2025-11-15; **Revised**: 2026-01-09

***Corresponding author**, **E-mail**: ldong90@seu.edu.cn.

ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 1,Jan. 2026,pp. 66－88
DOI:10. 16337/j. 1004-9037. 2026. 01. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

多无人机系统安全感知协同决策优化算法

李轶哲[1]，谢晨宇[1]，刘书鸣[2]，万子恒[1]，魏鑫锬[2]，董　璐[1]
(1. 东南大学网络空间安全学院，南京 211189；2. 东南大学计算机科学与工程学院，南京 211189)

摘　要：多无人机系统在动态环境下的协同决策面临安全性与鲁棒性的双重挑战。传统方法将安全机制与决策算法分离设计，难以保障系统在异常情况下的可靠运行。本文提出一种遵循“安全左移”与“设计安全”思想的协同策略优化框架(Adaptive security control with adversarial-resilient endogenous strategy，ASC-ARES)，通过状态建模与奖励塑形将拓扑控制、物理安全与能量管理等多层安全约束系统性嵌入深度强化学习(Deep reinforcement learning，DRL)决策过程，实现功能与安全的一体化设计。该框架首先扩展了深度确定性策略梯度(Deep deterministic policy gradient，DDPG)算法以适配混合动作空间，通过设计双头策略网络实现三维连续姿态与离散偏航角的联合优化。其次，框架融合了质心引导的双连通控制算法，赋予协同决策主动感知网络连通性的能力。最后，构建了以平均主观意见得分(Mean opinion score，MOS)为驱动的多目标自适应奖励机制，实现了用户体验质量(Quality of service，QoE)、网络双连通性、碰撞规避与能量效率的协同优化。实验结果表明，ASC-ARES框架具备优异的收敛特性与稳定性，其MOS波动率控制在0.36%，双连通成功率高达99.98%。对抗攻击实验显示，系统在遭受快速梯度符号法(Fast gradient sign method，FGSM)、投影梯度下降(Projected gradient descent，PGD)法及强噪声干扰($\epsilon=2.0$)后展现出优异的拓扑重构与状态恢复能力，干扰移除后的平均性能回升率超过80%。此外，消融实验证实了各安全组件的关键作用：拓扑控制模块将服务质量提升59%，排斥力机制则有效抑制了85%的碰撞风险。本研究为多无人机系统提供了一套平衡性能优化与内嵌安全保障的协同决策方案。

关键词：多无人机系统；深度强化学习；协同决策；设计安全；拓扑控制
中图分类号：TN92　　**文献标志码：**A
引用格式：李轶哲，谢晨宇，刘书鸣，等. 多无人机系统安全感知协同决策优化算法[J]. 数据采集与处理，2026，41(1)：66-88. LI Yizhe，XIE Chenyu，LIU Shuming，et al. A security-aware collaborative decision optimization algorithm for multi-UAV systems[J]. Journal of Data Acquisition and Processing，2026，41(1)：66-88.

引　　言

多无人机协同系统凭借其机动灵活、响应迅速、协同增效等优势，在军事侦察、应急救援、智能物流及通信中继等领域展现出广阔的应用前景[1-2]。相较于传统单机系统，多无人机协同系统依托分布式架构实现任务并行处理与资源动态配置，显著提升了任务执行效率与系统容错能力。然而，伴随应用场景日趋复杂，动态环境下协同决策的安全性与鲁棒性问题日益凸显，已成为制约该领域纵深发展的核心瓶颈。

从安全性角度审视，多无人机系统面临三重威胁。(1)网络拓扑结构脆弱。无人机高速机动导致网络拓扑持续演变，而现有拓扑控制方法存在固有缺陷：最小生成树方法受制于树形结构特性，难以规避

基金项目：国家自然科学基金(62576100)。
收稿日期：2025-11-15；**修订日期：**2026-01-09

单点故障风险[3];K近邻方法虽具备较高计算效率,却无法从理论层面确保双连通性[4]。关键节点一旦失效,极易引发网络分裂,致使协同任务中断。(2)外部攻击威胁严峻。多无人机系统易受恶意干扰与算法漏洞侵害,2023年某次军事演习中,因通信干扰导致无人机集群决策能力大幅下降,任务失败率高达40%[5],充分暴露了现有系统在对抗环境下的脆弱性。(3)多重物理约束交织。能量受限、碰撞规避、飞行边界等约束条件相互耦合,对决策算法的鲁棒性提出了更高要求。究其原因,上述问题源于传统方法沿用功能与安全分离的设计范式,安全模块作为附加组件叠加于功能模块之上,难以从根本上保障系统在异常工况下的稳定运行。

内生安全理论与智能决策技术的深度交叉为破解上述难题提供了新的契机。内生安全由邬江兴院士提出,其要旨在于通过架构级冗余与动态重构机制赋予系统与生俱来的免疫能力[6],使安全特性成为系统的固有属性而非外部附加。该理论蕴含两个核心思想:(1)设计安全,强调安全能力应在系统设计阶段即被纳入考量,而非作为事后补救措施;(2)安全左移,主张将安全机制前置嵌入开发流程的早期阶段,从源头消除安全隐患。深度强化学习(Deep reinforcement learning, DRL)则通过智能体与环境的持续交互习得最优策略[7],为动态环境下的协同决策提供了有力的方法论支撑。

现有内生安全研究主要聚焦于硬件与系统层面的拟态防御,通过多变体执行体的动态调度抵御未知攻击,已在Web服务器、工业控制系统等领域取得良好成效[8-9]。仝青等[9]的实验结果表明,采用异构执行体动态切换的拟态防御Web服务器可将攻击检测率提升60%。然而,将上述理念迁移至无人机协同决策算法层面,需正视两方面现实约束:一方面,无人机系统载荷与能量受限,难以承受硬件级多执行体冗余所带来的计算与存储开销;另一方面,高动态环境下网络拓扑持续演变,传统基于表决的异常检测机制难以有效区分正常拓扑变化与外部攻击行为。因此,在资源受限的无人机系统中直接复制硬件级方案既不经济也不可行,需探索算法层面的轻量化实现路径。

本文遵循“设计安全”与“安全左移”的思想,提出算法层面的安全一体化设计思路:通过在深度强化学习框架中系统性嵌入多层安全约束,使智能体在策略学习过程中自然习得安全行为模式,从而将安全特性从外部附加的“补丁”转化为决策策略的内在属性。具体而言,本文通过状态空间显式编码安全态势信息、奖励函数嵌入安全约束惩罚、学习参数依据性能状态自适应调整等机制,实现安全约束与功能优化的深度耦合。这种设计使系统在训练阶段即形成对安全边界的内在感知,在推理阶段无需额外安全监督模块即可自主规避风险,体现了“安全能力与生俱来”的设计理念。

就多无人机网络拓扑控制而言,拓扑控制是维系多无人机网络连通性与鲁棒性的关键技术[10]。纵观现有研究,相关方法可归纳为3类。基于距离的方法以K近邻算法为代表,计算复杂度为$O(n\log n)$,n为元素数量,但K值选取依赖先验经验,且理论上无法保证双连通性[4]。基于图论的方法以最小生成树为代表,边数虽最少(仅$n-1$条),但树形结构本身不具备双连通特性[3]。Bredin等[11]尝试通过增设中继节点实现双连通,但该方案在无人机系统中成本过高。基于优化的方法如Zhou等[12]提出的遗传算法拓扑优化方案,可将吞吐量提升20%~60%,但计算开销难以满足实时性要求。针对上述方法的局限,Li等[13]提出了质心引导的双连通拓扑控制算法,通过质心排序与双目标连接机制从设计层面确保网络双连通性,且无需人工预设参数。该算法为每个节点分配两个目标节点,确保任意节点均可经由两条独立路径连接至根节点。依据图论相关定理,双连通图等价于不存在割点的图[3],这意味着无单点故障、存在冗余路径且具备更强的生存能力。尽管如此,目前研究大多聚焦于拓扑控制算法本身,尚未深入探讨其与深度强化学习等决策算法的有机融合。

就深度强化学习在多无人机系统中的应用而言,深度强化学习为多无人机协同决策开辟了新的技术途径[7]。在基于值函数的方法中,深度网络算法(Deep Q-network, DQN)及其变体应用较为广泛。Liu等[14]采用Q-learning训练无人机动态调整轨迹,在静态场景下用户体验质量较传统K-means算法提

升30%；Wang等[15]将其改进为双网络架构，使任务成功率提升至85%，但离散动作空间的固有缺陷制约了控制精度。在基于策略梯度的方法中，深度确定性策略梯度（Deep deterministic policy gradient，DDPG）算法[16]与多智能体深度确定性策略梯度（Multi-agent deep deterministic policy gradient，MADDPG）算法[17]因可直接输出连续动作而备受关注；软演员-评论家（Soft actor-critic，SAC）算法[18]则引入最大熵机制增强了策略鲁棒性。此外，Li等[19]提出了基于深度强化学习的用户体验质量（Quality of service，QoE）驱动多无人机三维自适应部署策略，并结合安全共识控制机制提升系统鲁棒性。然而，现有深度强化学习方法仍存在明显不足。在动作空间设计方面，多数方法仅面向纯连续或纯离散动作空间，对于无人机飞行控制中常见的混合动作空间缺乏系统考量。在状态表示方面，现有方法通常仅包含无人机自身位置、速度及任务信息，缺乏对网络拓扑结构的显式建模。在安全约束集成方面，奖励函数设计往往偏重覆盖率、时延等任务指标，而忽视了拓扑连通性、碰撞规避、能量管理等安全约束的统一整合。例如，Liu等[20]提出的动态功率控制算法虽将误码率降至0.01%，却未将拓扑连通性纳入考量；Zhou等[12]虽关注用户体验质量优化，但同样未实现多层安全约束的统一建模。

综上所述，现有研究在应对复杂动态环境时仍存在以下局限性：一是功能与安全的架构性解耦，传统外挂式安全补丁难以实现决策过程中的风险主动感知与内嵌防御；二是安全约束建模的碎片化，缺乏将拓扑控制、物理避障与能量管理纳入统一状态空间的系统性框架，导致多维约束与任务收益间的耦合关系被割裂；三是评价体系的单一化，现有方法过度关注网络侧客观指标，缺乏以QoE为驱动的多目标协同优化机制，难以在动态环境中保障端到端的服务质量。针对这些挑战，本文的主要贡献如下：

（1）提出多无人机协同决策策略优化框架（Adaptive security controlwith adversarial-resilient endogenous strategy，ASC-ARES）。该框架将拓扑控制安全（质心引导的双连通性保障）、物理安全（碰撞规避与边界约束）及能量管理系统性集成至深度强化学习决策过程，通过状态编码显式建模安全态势、奖励塑形引导安全策略学习，消除网络单点故障隐患，实现功能与安全的协同设计。

（2）扩展深度确定性策略梯度算法以支持混合动作空间。针对无人机飞行控制的异质性需求，设计双头策略网络架构，连续分支优化俯仰角、横滚角、油门三维姿态控制，离散分支处理偏航角模式选择，通过非线性映射机制实现归一化控制量到物理执行指令的精确转换。

（3）构建面向用户体验质量的多目标自适应奖励机制。设计平均主观意见得分（Mean opinion score，MOS）驱动的分阶段优化策略，依据服务质量水平动态调整学习率与探索率衰减速度，在高性能区促进策略收敛、低性能区增强探索能力，实现快速收敛与稳定性能的平衡。

1 多无人机系统模型

本节构建多无人机协同通信与拓扑控制的系统模型，如图1所示，包括能量管理模型、通信信道模型、混合动作空间设计和服务质量评估机制，为ASC-ARES算法框架提供环境基础。

1.1 能量模型

考虑包含N架无人机的协同系统，服务区域为1 000 m × 1 000 m的二维平面，无人机飞行高度范围为$h\in[800,1\,500]$ m。每架无人机配备能量容量

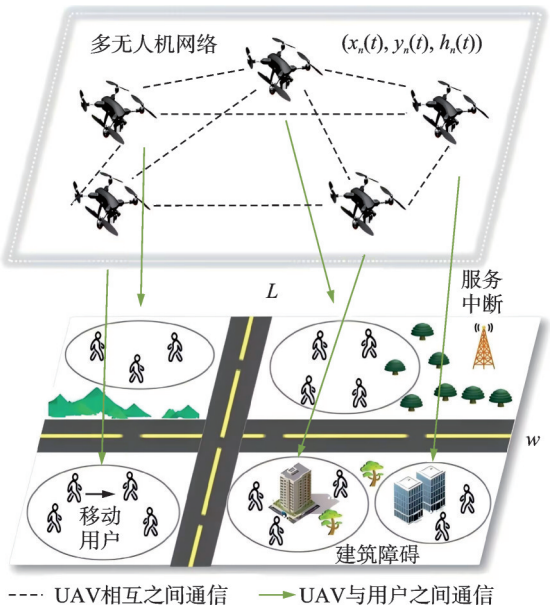


图1 多无人机协同通信系统架构示意图

Fig.1 Architecture of multi-UAV collaborative communication system

$E_{\max}=200$ J的电池系统,能量消耗包括悬停能耗、运动能耗和通信能耗3部分。悬停功率计算采用基于物理原理的诱导速度迭代模型。设无人机质量$m=2.0$ kg,螺旋桨半径$r=0.15$ m,旋翼数$n=4$,海平面空气密度$\rho=1.225$ kg/m^3,重力加速度$g=9.81$ m/s^2,悬停推力$T=mg$。诱导速度v_{ind}(螺旋桨诱导的向下气流速度)通过推力平衡方程迭代求解,即

$$v_{\text{ind}}=\frac{2T}{n\pi r^2\rho\sqrt{v_{\text{cruise}}^2+v_{\text{ind}}^2}} \tag{1}$$

式中$v_{\text{cruise}}=10$ m/s为巡航速度。悬停功率计算为

$$P_{\text{hover}}=\frac{v_{\text{ind}}T}{\eta} \tag{2}$$

式中功率效率$\eta=0.8$(考虑电机效率和空气动力学损失)。

运动能耗基于无人机位置变化计算。设无人机从位置$\boldsymbol{p}_t$移动到$\boldsymbol{p}_{t+1}$,位移距离$\Delta d=\|\boldsymbol{p}_{t+1}-\boldsymbol{p}_t\|$,飞行时间$\Delta t=\Delta d/v_{\text{cruise}}$,则运动能耗为$E_{\text{move}}=P_{\text{hover}}\cdot\Delta t/1\,000$。通信能耗根据服务用户数$N_{\text{users}}$动态调整:$E_{\text{comm}}=(P_{\text{transmit}}/N_{\text{users}})\times 0.1$,其中发射功率$P_{\text{transmit}}=20$ W,系数0.1为单时隙通信时间占比。系统集成太阳能收集机制,每时隙收集能量$E_{\text{solar}}=0.005\times E_{\max}=1$ J,确保持续运行能力。能量约束为$E_t\geqslant E_{\text{move}}+E_{\text{comm}}$,否则无人机停止移动并受到惩罚。

1.2 通信模型

系统采用概率性信道模型,综合考虑视距(Line-of-sight, LoS)和非视距(Non-line-of-sight, NLoS)传播特性。设载波频率$f_c=2$ GHz,光速$c=3\times10^8$ m/s,自由空间路径损耗常数$K_0=(4\pi f_c/c)^2$。无人机i位置为(x_i,y_i,h_i),地面用户k位置为$(x_k,y_k,0)$,二者之间的三维欧氏距离为

$$d_{ik}=\sqrt{(x_i-x_k)^2+(y_i-y_k)^2+h_i^2} \tag{3}$$

仰角定义为$\theta=\arcsin(h_i/d_{ik})$,表示用户观察无人机的角度。视距概率采用基于仰角的经验模型,考虑仰角对传播环境的影响,有

$$P_{\text{LoS}}=b_1\left|\frac{180\theta}{\pi}-\zeta\right|^{b_2} \tag{4}$$

式中:环境参数$b_1=0.36$、$b_2=0.21$为经验系数,$\zeta=15^\circ$为参考仰角。非视距概率$P_{\text{NLoS}}=1-P_{\text{LoS}}$。信道增益综合LoS和NLoS路径计算

$$G_{ik}=\frac{1}{K_0 d_{ik}^{\alpha}(P_{\text{LoS}}\mu_{\text{LoS}}+P_{\text{NLoS}}\mu_{\text{NLoS}})} \tag{5}$$

式中:路径损耗指数$\alpha=2$(自由空间传播),超额损耗因子$\mu_{\text{LoS}}=3$ dB和$\mu_{\text{NLoS}}=23$ dB分别表征LoS和NLoS路径的额外衰减。

用户通信速率基于香农容量公式,表达式为

$$R_{ik}=B_k\log_2\left(1+\frac{P_kG_{ik}}{\sigma^2}\right) \tag{6}$$

式中:$B_k=B_{\text{total}}/N_{\text{cluster}}$为用户分配带宽(总带宽$B_{\text{total}}=1$ MHz按簇均分),$P_k=P_{\max}/N_{\text{cluster}}$为用户分配功率(最大功率$P_{\max}=0.1$ W按簇内用户数均分),噪声功率$\sigma^2=B_k\cdot N_{\text{AWGN}}$,噪声功率谱密度$N_{\text{AWGN}}=10^{-20}$ W/Hz(对应-170 dBm/Hz)。

1.3 混合动作空间

多旋翼无人机飞行控制本质上具有异质性:姿态控制(俯仰、横滚、油门)要求连续精细调节以生成光滑飞行轨迹,而航向控制(偏航)因受限于转矩响应机制与机械约束,宜采用离散模式切换策略。然

而,传统深度确定性策略梯度算法天然假设动作空间连续,在处理此类混合决策问题时面临两难困境:强制连续化离散控制将引发策略震荡与训练不收敛,而完全离散化则丧失姿态控制所需的高精度与平滑性。鉴于上述挑战,本文构建混合动作空间 $\mathcal{A}=\mathcal{A}_{\mathrm{c}}\times\mathcal{A}_{\mathrm{d}}$,将连续与离散决策有机统一。连续子空间 $\boldsymbol{a}_{\mathrm{c}}=[a_{\mathrm{pitch}},a_{\mathrm{roll}},a_{\mathrm{throttle}}]\in[-1,1]^3$ 依次对应俯仰、横滚与油门的归一化控制量,离散子空间 $a_{\mathrm{d}}\in\{0,1,2\}$ 编码偏航的3种运动模态(逆时针旋转、航向保持、顺时针旋转)。该设计充分利用深度强化学习在连续控制上的精度优势,同时保留离散决策的稳定性。

动作映射机制将归一化控制量转换为物理执行指令。俯仰角 ϕ 与横滚角 ψ 通过双曲正切函数非线性映射至 $[-30^\circ,30^\circ]$ 区间:$\phi=30^\circ\cdot\tanh(a_{\mathrm{pitch}})$,$\psi=30^\circ\cdot\tanh(a_{\mathrm{roll}})$,该映射在零点附近具有高灵敏度而在边界处自然饱和,兼顾响应速度与系统稳定性。水平速度矢量由姿态角通过球面几何投影解算:$v_x=v_{\max}\sin\phi\cos\psi$,$v_y=v_{\max}\sin\phi\sin\psi$,其中巡航速度上界 $v_{\max}=15$ m/s符合旋翼气动效率峰值区间。垂直位移通过油门线性映射获得:$\Delta h=30\cdot a_{\mathrm{throttle}}$(单位:m),单步最大升降幅度30 m在保证充分机动性的同时维持安全裕度。偏航角速度直接由离散动作索引:$\omega_{\mathrm{yaw}}=\{-45^\circ/s,0,+45^\circ/s\}[a_{\mathrm{d}}]$,该角速度限制源于电机转矩能力与陀螺稳定性要求。

1.4 碰撞规避与区域限制

多无人机协同系统的物理安全需求体现为空间碰撞规避与任务区域约束两个维度。碰撞规避要求系统在动态环境中实时维护节点间安全间隔,防止相对运动导致的物理冲突;区域限制则确保无人机在执行任务时不偏离服务范围,同时保持合理的空间分布以优化覆盖性能。本文设计基于虚拟力场的物理安全约束机制。

碰撞规避机制。系统基于人工势场方法构建节点间排斥力模型,利用距离依赖的非线性势场引导无人机自主避碰。对于节点 i,其所受排斥力矢量定义为

$$\boldsymbol{F}_{\mathrm{repel},i}=\sum_{j\neq i}k_r\frac{\boldsymbol{p}_i-\boldsymbol{p}_j}{\|\boldsymbol{p}_i-\boldsymbol{p}_j\|^3}\tag{7}$$

式中 k_{r} 为排斥系数,随节点间距 $d_{ij}=\|\boldsymbol{p}_i-\boldsymbol{p}_j\|$ 减小呈指数增长。势场激活机制设计为:当 $d_{ij}<d_{\mathrm{safe}}=150$ m时势场激活并产生强排斥作用,当 $d_{ij}>R_{\mathrm{comm}}=650$ m时势场自然衰减至零。该分段设计确保近距避碰的高优先级,同时避免对远距离协同产生不必要的约束干扰。安全阈值150 m基于旋翼无人机机动性能与响应时延确定,通信半径650 m则对应网络连通性维持的最大节点间距。

区域限制机制。系统通过中心吸引力势场与边界约束联合实现无人机空间分布控制。中心吸引力势场定义为 $\boldsymbol{F}_{\mathrm{center},i}=k_{\mathrm{c}}(\boldsymbol{c}-\boldsymbol{p}_i)$,其中 $\boldsymbol{c}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{p}_i$ 为集群几何质心。吸引系数 k_{c} 根据偏离距离自适应调整:节点远离质心时吸引力增强,接近质心时减弱,形成柔性约束机制。水平维度的核心区域约束将可行位置限定于服务区域的5%~95%范围,该范围设定基于两方面考虑:下界(5%)防止节点过度聚集于中心导致边缘覆盖不足,上界(95%)防止节点滞留边界导致整体覆盖质量退化。垂直维度的高度约束基于空地通信模型推导获得:过低飞行高度(<800 m)因天线仰角限制导致覆盖范围受限并产生覆盖空洞,过高飞行高度(>1 500 m)因自由空间路径损耗激增导致信噪比不足以支持可靠通信。系统通过联合信噪比阈值(SNR≥10 dB)与几何仰角约束($\theta\in[30^\circ,60^\circ]$)计算可行高度区间,确保覆盖性能与通信质量的双重保障。

1.5 服务质量评估

系统采用MOS作为QoE指标,范围为1~5分(5分最优)。MOS计算基于端到端延迟 τ,延迟模型综合考虑往返时间、文件传输延迟和TCP慢启动效应。

延迟模型:设最大段大小 MSS=1 460×8=11 680 bits(标准以太网MTU),文件大小 FS=

$600\times1\,024\times8=4\,915\,200$ bit(600 KB),往返时间 RTT = 0.04 s(40 ms,典型移动网络)。传输控制协议慢启动阶段窗口增长轮数为

$$L=\min\left(\log_2\left(\frac{R_{ik}\cdot\mathrm{RTT}}{\mathrm{MSS}}+1\right)-1,\log_2\left(\frac{\mathrm{FS}}{2\mathrm{MSS}}+1\right)-1\right) \tag{8}$$

式中等号右边第1项为带宽时延积限制的窗口轮数,第2项为文件大小限制的窗口轮数,取二者最小值。总延迟包含3次握手时间、文件传输时间和慢启动损失,表达式为

$$\tau=3\mathrm{RTT}+\frac{\mathrm{FS}}{R_{ik}}+L\cdot\mathrm{RTT}-\frac{2\mathrm{MSS}(2^L-1)}{R_{ik}} \tag{9}$$

MOS映射:MOS值通过对数拟合延迟计算,表达式为

$$\mathrm{MOS}_{ik}=-C_1\ln(\tau_{ik})+C_2 \tag{10}$$

式中:拟合系数 $C_1=1.12$,$C_2=4.674\,6$,延迟下界 $\tau\geqslant\mathrm{RTT}$。无人机 i 服务簇 $\mathcal{U}_i$ 的平均主观意见得分为 $\mathrm{MOS}_i=\sum_{k\in\mathcal{U}_i}\mathrm{MOS}_{ik}$。系统总平均主观意见得分为所有无人机簇平均主观意见得分之和。

1.6 拓扑控制与双连通性

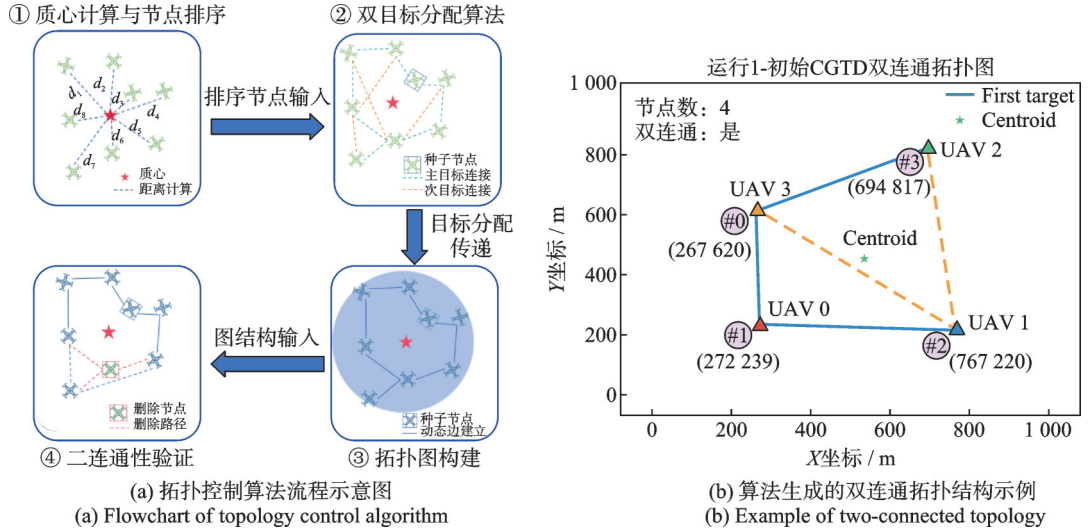
系统采用基于质心引导的双连通拓扑控制(Centroid-guided target-driven topology, CGTD)算法,确保网络无单点故障。双连通图的关键特性在于任意节点失效都不会导致网络分裂,为协同任务提供鲁棒性保障。拓扑控制算法流程为:

(1)质心计算与排序:计算所有无人机位置的几何质心 $\boldsymbol{c}_{\mathrm{geo}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{p}_i$,计算每个节点到质心距离 $d_i=\|\boldsymbol{p}_i-\boldsymbol{c}_{\mathrm{geo}}\|$,按距离升序排序生成节点序列 $\{v_0,v_1,\cdots,v_{N-1}\}$;

(2)目标分配:节点 v_0(距质心最近)作为根节点,节点 v_1 连接 v_0。对于节点 $v_i(i\geqslant2)$,分配两个目标节点以确保双连通性——第1目标 $t_1(i)$ 为前序节点 $\{v_0,v_1,\cdots,v_{i-1}\}$ 中距离最近者,第2目标 $t_2(i)$ 从 $t_1(i)$ 的邻居节点中选择距离最近者(若邻居集为空则选择次近前序节点);

(3)连接建立:通信半径 $R_{\mathrm{comm}}=650$ m,当UAV间距离 $\|\boldsymbol{p}_i-\boldsymbol{p}_{t_j(i)}\|\leqslant R_{\mathrm{comm}}$ 时建立边 $(v_i,v_{t_j(i)})$。

该算法从设计上保证每个节点通过两条独立路径连接到根节点,从理论上确保网络双连通性,如图2所示。



(a) 拓扑控制算法流程示意图
(a) Flowchart of topology control algorithm

(b) 算法生成的双连通拓扑结构示例
(b) Example of two-connected topology

图2 拓扑控制算法

Fig.2 Topology control algorithm

1.7 问题建模

多无人机协同覆盖优化问题具有典型的序贯决策特征：系统状态随无人机运动持续演化，每步决策需综合考虑服务质量、网络连通性、物理安全与能量约束等多维目标，且当前决策影响未来收益。鉴于问题的动态性与马尔可夫特性，本文采用马尔可夫决策过程（Markov decision process，MDP）框架建模协同决策问题，记为5元组$\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$。

状态空间$\mathcal{S} \subseteq \mathbf{R}^{56}$编码系统决策所需的完整观测信息。对于无人机$i$，其状态$\boldsymbol{s}_i$由3类信息构成：(1)自身三维位姿$\boldsymbol{p}_i = (x_i, y_i, h_i)^{\mathrm{T}} \in \mathbf{R}^3$，表征当前空间位置与服务能力；(2)服务用户簇位置$\boldsymbol{U}_i \in \mathbf{R}^{2K_{\max}}$，编码最多$K_{\max} = 25$个地面用户的二维坐标，不足部分以零填充；(3)航向角度编码$\boldsymbol{\phi}_i = (\sin\psi_i, \cos\psi_i, \psi_i)^{\prime} \in \mathbf{R}^{\mathrm{T}}$，通过三角函数解决角度周期性导致的学习困难。状态向量由上述分量组合而成，表达式为

$$\boldsymbol{s}_i = [\, \boldsymbol{p}_i^{\mathrm{T}}, \boldsymbol{U}_i^{\mathrm{T}}, \boldsymbol{\phi}_i^{\mathrm{T}} \,]^{\mathrm{T}} \in \mathbf{R}^{56} \tag{11}$$

该状态表示在保证马尔可夫性的前提下，兼顾维度精简与信息完备性。系统采用混合动作空间$\mathcal{A} = \mathcal{A}_{\mathrm{c}} \times \mathcal{A}_{\mathrm{d}}$，连续子空间$\mathcal{A}_{\mathrm{c}} = [-1, 1]^3$控制姿态调节，离散子空间$\mathcal{A}_{\mathrm{d}} = \{0, 1, 2\}$控制航向模式。转移概率$\mathcal{P}(\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, \boldsymbol{a}_t)$由运动学模型、能量约束与用户移动模型共同决定，满足马尔可夫性假设。

奖励函数$\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbf{R}$采用多目标加权机制，将服务质量优化与安全约束统一于单一标量信号

$$R(\boldsymbol{s}, \boldsymbol{a}) = w_{\mathrm{mos}} \Delta\mathrm{MOS} + R_{\mathrm{perf}} + R_{\mathrm{stab}} + R_{\mathrm{maint}} + R_{\mathrm{conn}} + R_{\mathrm{coll}} + R_{\mathrm{alt}} - w_{\mathrm{e}} E_{\mathrm{cost}} \tag{12}$$

式中等号右边前4项构成性能优化子目标，后4项构成安全与能效约束。关键设计包括：(1)MOS增量项$w_{\mathrm{mos}}\Delta\mathrm{MOS}$通过自适应权重（低性能480、中性能400、高性能240）实现分阶段优化策略；(2)拓扑连通性项R_{conn}采用强激励机制（±100）确保网络鲁棒性；(3)碰撞与高度约束项R_{coll}、R_{alt}通过硬惩罚（−50）强制物理安全；(4)能量项$w_{\mathrm{e}} E_{\mathrm{cost}}$引导能效优化但不主导决策。为稳定训练，系统采用指数移动平均平滑即时奖励为

$$R_{\mathrm{smooth}}(t) = \alpha_{\mathrm{ema}} R_{\mathrm{smooth}}(t-1) + (1 - \alpha_{\mathrm{ema}}) R_{\mathrm{raw}}(t) \tag{13}$$

最终反馈为$0.9R_{\mathrm{raw}}(t) + 0.1R_{\mathrm{smooth}}(t)$，平衡响应速度与方差抑制（平滑系数$\alpha_{\mathrm{ema}} = 0.05$）。

学习最优策略$\pi^*: \mathcal{S} \times \mathcal{A}$以最大化期望累积折扣回报，即

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(\boldsymbol{s}_t, \boldsymbol{a}_t) \right] \tag{14}$$

式中：$\gamma = 0.99$为折扣因子，$\tau = (\boldsymbol{s}_0, \boldsymbol{a}_0, r_0, \boldsymbol{s}_1, \boldsymbol{a}_1, r_1, \cdots)$为策略$\pi$与环境交互产生的状态-动作-奖励轨迹。

2 ASC-ARES算法框架

2.1 算法设计理念

ASC-ARES算法框架遵循“设计安全”的理念，通过将安全机制前置内嵌于决策闭环，实现功能优化与安全保障的一体化设计，如图3所示。其核心特性体现在以下3个维度：

(1)MOS驱动的自适应优化机制：系统构建了以MOS为核心的性能反馈闭环，实现了性能的动态调控。该机制依据QoE实时水平，在不同性能区间实施差异化的超参数调度策略：在低性能区维持高学习率并减缓探索噪声衰减，以强化对安全策略空间的全局搜索；在高性能区降低学习率并加速噪声收敛，以促进策略在安全边界内的精细寻优。这种基于性能感知的自适应调节，有效平衡了强化学习中探索与利用的矛盾，确保系统在不突破安全底线的前提下实现性能最大化。

(2)混合动作空间强化学习内核：针对无人机飞行控制中“连续姿态调节”与“离散航向决策”并存的特性，框架扩展了DDPG算法。策略网络采用双分支架构，分别输出三维连续控制量与离散动作概

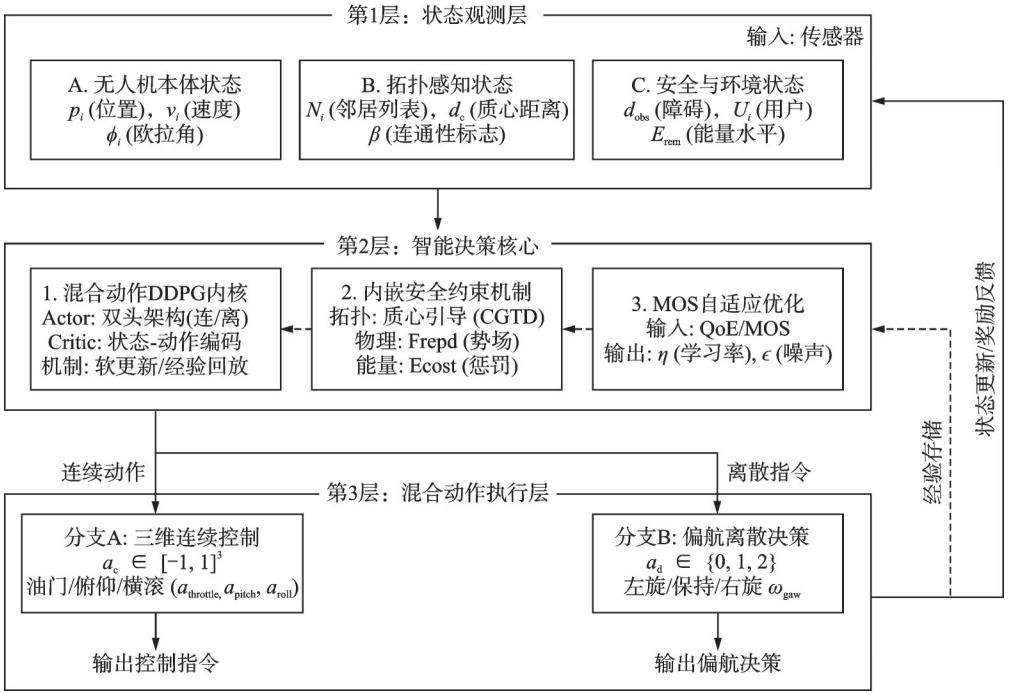


图3 ASC-ARES算法框架整体架构

Fig.3 Overall architecture of ASC-ARES algorithm framework

率，价值网络则通过联合编码机制实现状态-混合动作对的统一价值估计。同时，结合离线学习与经验回放机制提升样本效率，并引入目标网络软更新技术以抑制训练过程中的Q值震荡。

(3)内嵌式多层安全约束集成：系统将拓扑连通、物理安全与能量管理3类硬约束系统性地映射为状态空间特征与奖励函数分量，实现多维约束的深度嵌入。具体而言：拓扑层融合质心引导算法，将双连通性校验结果内化为决策状态；物理层引入人工势场法，通过相对位置编码与非线性排斥力奖励实现主动避障与区域保持；能量层基于物理能耗模型引导高能效飞行决策。这3层约束协同作用，将原本离散的安全规则转化为智能体内生的策略偏好。

2.2 深度神经网络架构

2.2.1 策略网络结构

策略网络采用共享特征提取与双分支输出的架构设计。特征提取模块由3层全连接网络构成(256→128→64)，逐层压缩状态表示，各层配置LayerNorm以稳定训练过程，并采用ReLU激活函数引入非线性变换。输出模块分为2个独立分支：连续动作分支生成三维姿态控制指令(油门、俯仰、横滚)，经tanh激活函数映射至[−1, 1]归一化区间；离散动作分支输出三维logits向量，经Softmax归一化得到偏航模式的概率分布。网络参数采用Xavier初始化，训练过程引入Dropout机制抑制过拟合。

2.2.2 价值网络结构

价值网络采用双路编码与特征融合的架构实现Q值估计。状态编码路径对状态向量依次施加128维全连接变换与LayerNorm归一化，得到状态特征表示；动作编码路径首先将连续动作(三维)与离散动作的独热编码(三维)组合为6维向量，再经64维全连接层提取动作特征。两路特征拼接后送入融合模块，该模块由两层全连接网络(128→64→1)构成，最终输出标量Q值。输出层采用小增益初始化策略(gain=0.1)，缓解训练初期的价值过估计问题。

2.2.3 目标网络机制

系统为每个无人机维护策略网络与价值网络的目标副本$\mu_{\theta'}$和$Q_{\phi'}$,用于生成稳定的时序差分(Temporal difference, TD)目标。目标网络参数通过软更新机制追踪主网络:$\theta' \leftarrow \tau\theta + (1-\tau)\theta'$,其中软更新系数$\tau = 0.005$控制追踪速度。该机制有效抑制目标值振荡,提升训练稳定性。4无人机系统共维护16个神经网络(每架4个:策略主网络、策略目标网络、价值主网络、价值目标网络)。

2.3 训练机制

ASC-ARES框架采用离线策略学习范式构建完整训练流程,通过经验回放、批量学习与目标网络机制协同作用,实现稳定高效的策略优化。

2.3.1 经验回放与批量学习

系统采用容量为1 000的经验回放缓冲区存储历史交互经验。当缓冲区样本数达到批量规模$B = 32$时,随机采样一批经验元组$\{(\boldsymbol{s}^{(k)}, \boldsymbol{a}^{(k)}, r^{(k)}, \boldsymbol{s}'^{(k)})\}_{k=1}^{B}$进行网络更新。

价值网络更新通过最小化时序差分误差更新价值网络参数,即

$$\mathcal{L}_Q = \frac{1}{B}\sum_{k=1}^{B}(Q_\phi(\boldsymbol{s}^{(k)}, \boldsymbol{a}^{(k)}) - y^{(k)})^2 \tag{15}$$

式中TD目标$y^{(k)}$由目标网络与折扣因子$\gamma = 0.99$计算得到。采用学习率$\eta_Q = 4 \times 10^{-4}$,并对梯度施加阈值裁剪(上限0.8)防止梯度爆炸。

策略网络更新通过确定性策略梯度$\nabla_\theta \mathcal{J}_\mu = \mathbb{E}[\nabla_{\boldsymbol{a}} Q_\phi \nabla_\theta \mu_\theta]$最大化期望累积回报。混合动作空间下,连续动作梯度经链式法则反向传播,离散动作采用概率加权期望形式。策略网络学习率设为$\eta_A = 2 \times 10^{-4}$(为价值网络的一半),采用Adam优化器,每次更新后对目标网络执行软更新。

2.3.2 噪声探索策略

确定性策略框架需通过显式噪声机制促进环境探索。系统采用自适应噪声注入策略,噪声强度随训练进展动态衰减。连续动作空间采用加性高斯噪声:$\tilde{\boldsymbol{a}}_c = \boldsymbol{a}_c + \epsilon$,其中$\epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_3)$,标准差$\sigma$从初始值0.2逐步衰减至0.01。离散动作空间采用ϵ-softmax混合策略:以概率0.5ϵ执行随机选择,否则依据网络输出概率分布采样。探索率ϵ在训练过程中从0.3衰减至0.05,衰减速度由平均主观意见得分水平驱动:高性能区采用较快衰减(因子0.997 5)促进策略稳定,低性能区采用较慢衰减(因子0.999 5)增强探索能力,实现探索与利用的自适应平衡。

2.3.3 自适应学习率调整

系统根据服务质量水平动态调整学习率:$\eta_t = \eta_{\text{base}} \cdot \lambda_\eta(\text{MOS}_t)$,其中$\lambda_\eta(\cdot)$为平均主观意见得分驱动的自适应系数函数,$\text{MOS}_t$表示$t$时刻的平均主观意见得分。高性能区降低学习率(系数0.7)以稳定策略,低性能区保持基准学习率(系数1.0)以持续优化。价值网络学习率(4×10^{-4})设为策略网络(2×10^{-4})的2倍,加速价值函数收敛并为策略更新提供准确的梯度信号。

2.4 多层安全机制嵌入

系统将拓扑控制安全、物理安全和能量管理系统性集成到深度强化学习决策过程,通过奖励塑形、状态感知和硬约束的协同作用,实现安全与性能的一体化优化。

2.4.1 拓扑控制层集成

质心引导的拓扑控制算法深度集成到深度强化学习训练循环,每个时隙执行目标分配与双连通性验证。集成机制通过3个层面协同作用:(1)状态感知层面,拓扑算法输出的双连通性验证结果通过状态编码影响策略学习;(2)奖励塑形层面,双连通状态获得+100奖励,非连通状态受到−100惩罚,形成强对比的引导信号;(3)决策闭环层面,拓扑控制与策略学习形成闭环反馈——无人机位置更新→拓扑

重计算→连通性验证→奖励反馈→价值估计更新→策略梯度调整,确保网络鲁棒性与任务性能的统一优化。

2.4.2 物理安全层集成

物理安全约束通过双路径机制融入深度强化学习决策过程。状态建模路径将排斥力矢量、中心吸引力、边界距离及碰撞风险指标等物理约束信息显式编码为观测维度,使策略网络直接感知安全态势;奖励塑形路径将碰撞惩罚、边界惩罚及安全间隔奖励等嵌入即时奖励函数,通过价值函数传播引导长期策略优化。

碰撞规避采用软硬约束结合的双重机制:当无人机间距小于50 m时,奖励函数施加−50惩罚作为软约束;当间距小于150 m时,势场排斥力直接修正位置作为硬约束,形成预防性避障机制。动态高度约束根据信噪比要求与仰角约束实时计算可行范围,违反约束时施加−50惩罚并强制修正飞行高度。服务区域约束通过中心吸引力与核心区域限制防止无人机偏离覆盖目标,确保持续有效的服务能力。上述机制通过奖励函数的安全约束项(R_{coll}, R_{alt})协同作用,使智能体在训练阶段内化物理安全规则,执行时无需外部安全监督模块。

2.4.3 能量管理层集成

能量管理基于精确物理模型通过奖励引导机制实现能效优化。系统为每架无人机维护能量状态,实时计算能量消耗与太阳能收集量。奖励函数中的能量消耗惩罚项$w_{\text{e}}E_{\text{cost}}$将能量效率纳入优化目标,权重设置确保能量因素参与策略学习但不主导决策方向。当能量储备不足时,系统强制无人机停止移动并施加额外惩罚,促使策略学习合理的能量分配模式。该机制通过精确物理建模、持续太阳能充电与奖励引导的协同作用,在保证任务性能的前提下引导策略向能量高效方向演化。

2.5 完整算法流程

ASC-ARES算法框架的完整流程如算法1所示。算法包含3个嵌套循环:外层为运行次数循环(多次独立实验),中层为时隙循环(每次运行的时间步),内层为智能体循环(多无人机并行决策)。关键步骤包括:(1)环境初始化与K-means聚类进行用户簇划分和无人机位置初始化;(2)每时隙执行拓扑控制模块获取双连通状态;(3)各智能体基于局部状态选择动作并与环境交互;(4)经验存储至回放缓冲区;(5)随机采样批量经验更新策略网络和C价值网络;(6)目标网络软更新;(7)基于平均主观意见得分性能动态调整学习率和探索率。

算法 1 ASC-ARES 算法框架

Require:无人机数量N,用户数量M,时隙数$T_{\max}$,运行次数N_{runs}

Require:学习率η_A, η_Q,折扣因子γ,软更新系数τ,批量大小B

Ensure: 最优策略网络$\{\mu_{\theta_i^*}\}_{i=1}^{N}$

(1) for run = 1 to N_{runs} do

(2) 初始化策略网络$\{\mu_{\theta_i}\}$,价值网络$\{Q_{\phi i}\}$及其目标网络

(3) 初始化经验回放缓冲区$\mathcal{D}\leftarrow\varnothing$,探索率$\epsilon\leftarrow\epsilon_0$,噪声$\sigma\leftarrow\sigma_0$

(4) 生成用户位置并K-means聚类划分用户簇$\{U_i\}_{i=1}^{N}$,初始化UAV位置

(5) for t = 1 to $T_{\max}$ do

(6) 执行二连通拓扑控制,验证双连通性β

(7) for UAV i = 1 to N do

(8) 观测状态$\boldsymbol{s}_i$,Actor前向传播得$\boldsymbol{a}_{\text{c},i}$, $\mathbf{z}_{\text{d}}$, i

(9) 连续动作加噪声$\tilde{a}_{\text{c},i}\leftarrow\boldsymbol{a}_{\text{c},i}+N(\mathbf{0},\sigma^2\boldsymbol{I}_3)$,离散动作采样$a_{\text{d},i}$

(10) 执行动作，观测新状态 s_i' 和奖励 r_i

(11) end for

(12) 存储联合经验至 $\mathcal{D}$

(13) if $|\mathcal{D}| \geqslant B$ then

(14) 从 $\mathcal{D}$ 随机采样批量 $\mathcal{B}$

(15) for UAV $i = 1$ to N do

(16) 计算时序差分目标：$y^k = \alpha_\tau r_i^k + \gamma Q\phi_i'(s_i'^{(k)}, \mu\theta_i'(s_i'^{(k)}))$

(17) 更新价值：$\phi_i \leftarrow \phi_i - \eta_Q \nabla\phi_i \mathcal{L}_Q$

(18) 更新策略：$\theta_i \leftarrow \theta_i + \eta_A \nabla\theta_i \mathcal{J}_\mu$

(19) 软更新目标网络：$\theta_i' \leftarrow \tau\theta_i + (1-\tau)\theta_i'$, $\phi_i' \leftarrow \tau\phi_i + (1-\tau)\phi_i'$

(20) end for

(21) end if

(22) 计算 $\mathrm{MOS}_{\mathrm{total}}$，基于平均主观意见得分自适应调整 η_A, η_Q, ϵ, σ

(23) end for

(24) end for

(25) return 训练好的策略网络 $\{\mu_{\theta_i^*}\}_{i=1}^N$ 所有实验中保持一致。无人机位置使用 K-means 算法进行初始聚类

3 仿真实验与分析

本节通过4组层次化实验系统验证ASC-ARES框架的核心贡献：稳定性验证实验评估算法收敛特性与对初始条件的鲁棒性；深度强化学习内核对比实验论证深度确定性策略梯度算法的适配性；拓扑控制算法对比实验通过基线性能与可扩展性双维度分析揭示质心引导双连通拓扑控制算法的理论保证优势；两组消融实验分别量化拓扑控制模块和物理安全机制对系统性能的贡献。实验设计遵循控制变量原则，每组实验仅改变单一因素，确保因果推断的有效性。

3.1 实验设置与参数配置

所有实验采用统一配置确保公平性：服务区域1 000 m × 1 000 m，4架无人机服务100个地面用户，最大能量200 J，飞行高度范围800～1 500 m，通信半径650 m，每次运行1 000时隙。深度强化学习参数为：策略网络学习率 $\eta_A = 2 \times 10^{-4}$，价值网络学习率 $\eta_Q = 4 \times 10^{-4}$，折扣因子 $\gamma = 0.99$，批量大小 $B = 32$，缓冲区容量1 000。用户位置在所有实验中保持一致。无人机位置使用 K-means算法进行初始聚类。具体设置详见附录A。

3.2 算法稳定性验证

稳定性验证通过5次独立训练评估ASC-ARES算法的鲁棒性和收敛一致性。每次训练执行1 000个时隙，采用不同随机种子初始化网络参数和探索噪声，观测MOS演化和算法收敛行为。

图4展示了5次独立训练的总MOS演化曲线。实验结果揭示了算法的3个学习阶段特性。初始探索阶段(0～

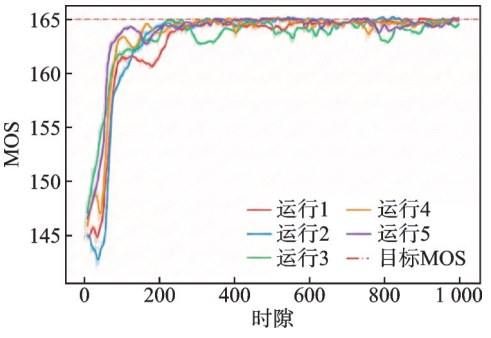


图4 ASC-ARES算法MOS性能演化曲线(5次独立运行)

Fig.4 MOS performance evolution curves of ASC-ARES algorithm (five independent runs)

50时隙):5次独立训练的初始MOS值分布于143～148,反映了随机初始化带来的合理差异。个别训练(如运行2)在前50时隙出现短暂下降至143的探索行为,随后迅速回升,体现了算法在早期探索过程中的自适应调整机制。实验表明,不同随机种子导致的初始网络参数差异对后续收敛过程无显著影响。快速学习阶段(50～200时隙):所有训练均表现出显著的学习效率,MOS值以近似线性的速率增长。各训练的收敛速度存在差异:最快收敛于80时隙达到目标值(运行5),最慢收敛于180时隙完成(运行3),平均收敛时间为120±40时隙,约占总训练时长的12%。尽管学习轨迹存在差异,但所有训练最终均收敛至相同目标,验证了算法对初始条件的鲁棒性。稳定收敛阶段(200～1 000时隙):200时隙后,所有训练均稳定维持在目标值165附近,平均MOS值为165.2±0.6,变异系数仅为0.36%。稳定阶段MOS值在164～166区间内小幅波动,波动幅度控制在±1以内。5条训练曲线在后800时隙内高度一致,未出现性能退化或异常波动现象。个别时隙的性能轻微下降均能快速恢复,表明算法具备良好的自适应调整能力。

综合性能评估表明,ASC-ARES算法在5次独立实验中实现了100%的收敛成功率,平均收敛时间约为120时隙,稳定阶段的MOS波动率为0.61%(标准差/均值),充分验证了算法的稳定性与鲁棒性。

3.3 深度强化学习内核对比实验

为选择最优的深度强化学习内核,系统对比了5种主流算法:DDPG[21]、近端策略优化(Proximal policy optimization, PPO)[22]、SAC[18]、演员-评论家(Actor-critic, AC)[23]和DQN[24]。实验在保持拓扑控制、安全约束机制、奖励函数设计等模块完全一致的前提下进行,确保对比公平性。所有算法采用统一的256-128-64网络架构,运行1 000时隙。实验结果如图5所示,5种算法呈现显著的性能差异。DDPG与PPO表现出最快的收敛速度,MOS值从初始143迅速上升,分别在80～100时隙和100～120时隙收敛至163并保持稳定。AC算法性能居中,在400～600时隙达到MOS值163,但学习过程存在明显的停滞平台期,且后期出现性能下降,稳定性不足。SAC和DQN算法性能较差,分别稳定在154～155和153。DQN因动作离散化导致控制精度受限,难以实现无人机位置的精细调节;SAC的熵正则化机制引入过度随机性,在确定性控制场景下降低了收敛效率。

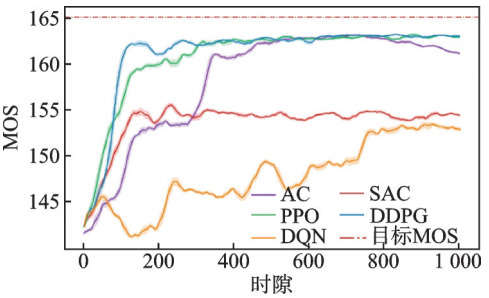


图5 深度强化学习算法MOS性能对比(1 000时隙)

Fig.5 MOS performance comparison of deep reinforcement learning algorithms (1 000 time slots)

长期稳定性方面,DDPG收敛后MOS值维持在163～163.5,波动幅度控制在±0.5以内,优于PPO和SAC。DDPG的优势源于以下核心机制:(1)确定性策略直接输出动作均值,避免了随机策略的高方差问题;(2)离线策略学习通过经验回放机制提升样本利用效率,支持多次迭代学习;(3)目标网络软更新机制提供稳定的时序差分学习目标,有效抑制了训练震荡;(4)本文扩展的混合动作空间机制支持连续控制(俯仰角、横滚角、油门)与离散决策(偏航模式)的统一建模,充分发挥了确定性策略在连续动作空间中的优势。

综合考虑收敛速度、最终性能和长期稳定性,DDPG成为本文ASC-ARES框架深度强化学习内核的最优选择,有效保障了多无人机协同决策的效率与可靠性。

3.4 拓扑控制算法对比实验

为了全面评估ASC-ARES框架中拓扑控制模块的性能,本节设计了系统性对比实验,将本框架的CGTD算法与4种经典拓扑控制方法进行比较:K-近邻(K-nearest neighbors,KNN)算法,$K=2$;最小生

成树(Minimum spanning tree, MST)算法;全连接策略以及无拓扑控制基线。实验从基线性能和可扩展性两个维度,对各算法在双连通性保障、网络开销、执行效率及通信代价等方面的表现进行了量化评估。

3.4.1 基线性能对比

基线实验固定无人机数量$N=4$,在通信半径$R\in\{650,700,750,800,850,900\}$ m下评估算法性能,每组配置独立运行50次。图6展示了双连通性、连通性、执行时间和网络开销4个维度的对比结果。双连通性:KNN算法在所有通信半径下保持100%双连通率,有效消除了单点故障风险。CGTD在650~700 m半径下双连通率约98%~99%(受限于通信半径不足以建立双目标连接),大于等于750 m时达到100%。全连接策略始终保持100%。MST由于其树状拓扑结构特性,无法提供双连通保障(双连通率为0%)。无拓扑控制组几乎不具备双连通能力(约1%)。

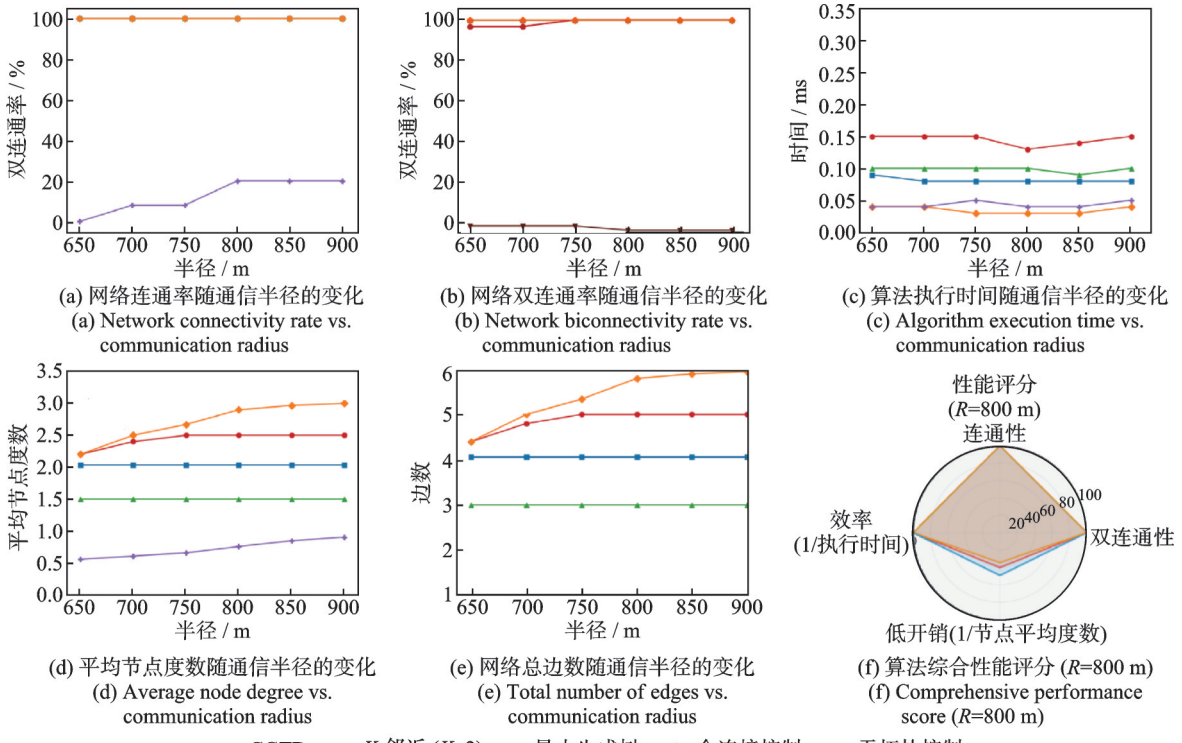


图6 拓扑控制算法基线性能综合对比(4架无人机,半径650~900 m)

Fig.6 Comprehensive baseline performance comparison of topology control algorithms (four UAVs, radius 650—900 m)

连通性:CGTD、KNN、MST及全连接策略在所有测试半径下均维持100%网络连通性。无拓扑控制组的连通率高度依赖通信半径,仅从650 m的5%提升至900 m的约20%,无法满足协同通信需求。

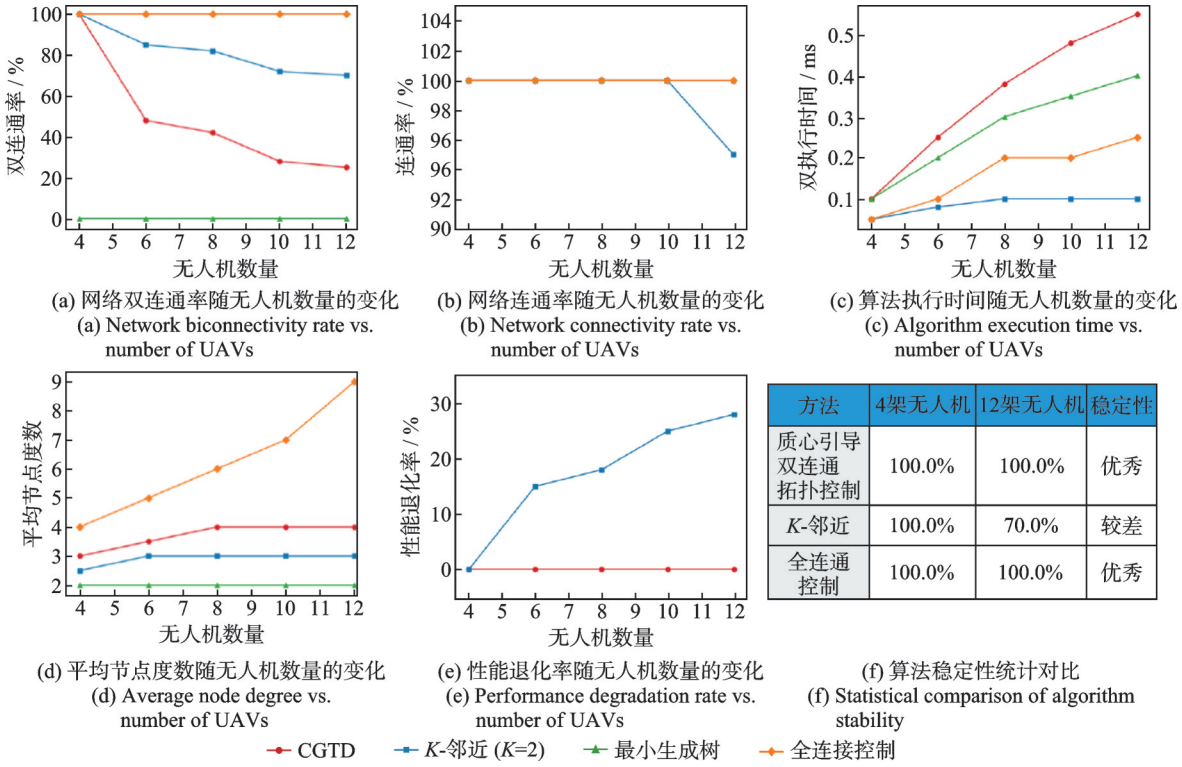
执行时间与通信开销:全连接策略执行效率最高(约0.047 ms),仅涉及距离判定。KNN次之(0.081 ms),需K次最近邻搜索。MST耗时中等(0.116 ms)。CGTD算法耗时相对较长(0.204 ms,为KNN的2.5倍),主要源于质心计算与双目标优选过程引入的计算负载。值得注意的是,尽管CGTD依赖全网位置信息,但在本实验设定的中小规模集群下,利用现有高带宽数据链广播位置状态的通信延迟处于毫秒级,工程上完全可行且不会引发信令拥塞。

网络开销(以$R=800$ m时的平均节点度衡量):MST节点度最低(1.50),但不具备双连通性。KNN节点度为2.03,在保证双连通性的前提下实现了最低开销(归因于$K=2$的固定约束)。CGTD节点度为2.43,虽比KNN高20%但仍处于合理范围。全连接策略节点度高达2.71,通信冗余开销最大。

综合效能评估:雷达图分析显示,在4架无人机的小规模场景下,KNN算法在双连通性、连通性、执行效率及低开销4个维度表现最为均衡。这主要归因于$K=2$的参数设置恰好满足4节点双连通的拓扑下界,且初始节点分布较为均匀,使得KNN达到了特定参数下的局部最优适配。然而,该优异性能在网络规模动态变化下的鲁棒性尚需进一步验证。

3.4.2 可扩展性实验

为验证算法的规模适应能力,可扩展性实验固定通信半径$R=800$ m,遍历无人机数量$N\in\{4,6,8,10,12\}$,评估各算法性能随网络规模扩展的演化趋势,结果如图7所示。



方法	4架无人机	12架无人机	稳定性
质心引导双连通拓扑控制	100.0%	100.0%	优秀
K-邻近	100.0%	70.0%	较差
全连通控制	100.0%	100.0%	优秀

图7 拓扑控制算法可扩展性对比分析(半径800 m,4~12 UAVs)

Fig.7 Scalability comparison of topology control algorithms (radius 800 m, 4—12 UAVs)

双连通性随规模变化:CGTD算法与全连接策略在所有测试规模(4~12架无人机)下均保持100%双连通率,表现出优异的可扩展性。相反,KNN的双连通率随节点规模增加呈显著退化趋势:从4架无人机时的100%,降至12架时的70%,性能退化率达30%。该现象揭示了KNN算法的参数敏感性缺陷:固定连接数$K=2$在大规模网络中无法提供足够的冗余路径以保障双连通性;若增加K值虽可改善大规模场景性能,却会显著增加小规模场景的冗余开销。MST在所有规模下双连通率均为0%。

时间复杂度的演化:CGTD算法的执行时间呈$O(n\log n)$增长(从0.20 ms增至0.55 ms)。KNN虽绝对耗时最小,但增长趋势与CGTD相近。全连接策略的时间复杂度为$O(n^2)$,从0.05 ms激增至0.25 ms。尽管CGTD引入了约0.12 ms的额外计算开销,但在无人机系统100 ms的控制周期内可忽略

不计，以微小的计算增量换取理论上的双连通性保障具有显著的工程价值。此外，针对未来大规模集群可能面临的通信瓶颈，本算法架构在理论上支持引入分布式平均一致性算法（Distributed average consensus, DAC），通过邻居间迭代交换信息估算全局质心，从而消除对全网广播的依赖，确保算法在更大规模网络中的扩展能力。

网络开销的增长趋势：CGTD算法的平均节点度从2.43增至3.50，增长约44%，增长速率处于可控范围。全连接策略从2.71急剧上升至11.00，增幅达306%，在大规模场景下将产生过高的通信负载。KNN由于算法约束，节点度恒定为2.00。

双连通性退化率：KNN算法的性能退化曲线清晰地反映了其结构性局限。其在小规模场景下的优秀表现高度依赖于“恰好”的K值选择与理想的节点分布；一旦网络规模变化或拓扑结构复杂化，其性能将显著下降。

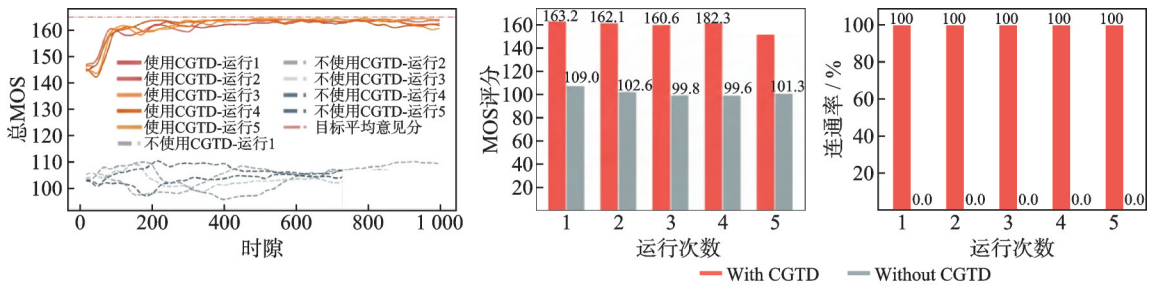
综上所述，ASC-ARES框架中的CGTD展现出更强的鲁棒性与通用性，是动态多无人机网络场景下的最优拓扑控制解决方案。

3.5 拓扑控制消融实验

拓扑控制消融实验通过对比有无拓扑控制模块的系统性能，量化拓扑控制对整体服务质量和网络连通性的贡献。实验设置两组对照：拓扑控制组集成完整的CGTD算法，每时隙执行质心引导的目标分配和双连通性验证；空白组移除拓扑控制模块，仅保留深度强化学习决策和物理安全机制。两组实验各运行5次，每次1 000时隙。

3.5.1 MOS性能对比

图8(a)展示了两组实验的MOS演化曲线。拓扑控制组（实线）在初始探索阶段（0～50时隙），MOS从145快速提升至160；快速学习阶段（50～200时隙），以陡峭斜率攀升至目标值165附近，稳定收敛阶段（200～1 000时隙），MOS稳定维持在162.2 ± 1.5，平均收敛时间为120时隙。相比之下，空白组（虚线）呈现显著的性能退化：初始MOS就因为无法通信而较低，并且学习过程极为缓慢且不稳定。在前200时隙内，MOS徘徊在100～108区间，部分运行甚至出现性能崩溃（运行3在时隙50～150之间MOS降至20，疑似网络完全失连导致无人机无法服务用户）。200时隙后，空白组的MOS逐步恢复并稳定在102.0 ± 5.2，但波动幅度是拓扑控制组的3.5倍，展现出极低的稳定性。



(a) MOS性能对比
(a) MOS performance comparison

(b) 综合性能对比
(b) Comprehensive performance comparison

图8 拓扑控制消融实验结果

Fig.8 Results of topology control ablation experiments

最终性能对比显示，拓扑控制组的平均MOS为162.2（最后50步均值），空白组仅为102.0，拓扑控制模块使系统服务质量提升59.0%。该提升主要源于：(1)双连通拓扑保障了可靠的通信链路，避免因单点故障导致的服务中断；(2)质心引导的部署优化了无人机覆盖范围，减少了用户端到端延迟；(3)拓

扑连通性作为奖励信号引导策略学习,促使网络学习维护鲁棒拓扑的协同移动模式。

3.5.2 连通性与综合性能分析

图8(b)展示了两组实验的综合性能对比。连通率方面,拓扑控制组在5次运行中均保持100%连通率,展现完美的网络连通性保障能力。空白组的连通率为0%,网络频繁出现分割现象,部分无人机因距离过远无法建立通信链路,导致用户服务中断和MOS骤降。最终MOS对比显示,拓扑控制组的5次运行MOS值集中在160.6~163.6(平均162.2),变异系数0.9%。空白组的MOS值分布在99.6~108.0(平均102.0),变异系数5.1%,是拓扑控制组的5.7倍,表明无拓扑控制时系统性能高度不稳定。

实验结果表明,拓扑控制模块是ASC-ARES框架的关键组件,对系统服务质量和网络连通性具有决定性作用。移除拓扑控制后,尽管深度强化学习算法仍可驱动无人机移动,但缺乏全局协调导致网络频繁失联,服务质量退化59%。该消融实验充分验证了"设计安全"理念的有效性——将安全机制前置嵌入决策过程,实现功能与安全的一体化优化。

3.6 物理安全机制消融实验

物理安全机制消融实验评估中心吸引力和无人机间排斥力对碰撞规避的贡献。如图9所示,实验设置3个消融配置:去除中心吸引力、去除排斥力、去除两种机制,基线结果与图4一致(MOS为165.2±0.6)。每个配置运行5次,每次1 000时隙。

3.6.1 去除中心吸引力实验

去除中心吸引力实验结果如图9(a)所示。实验结果显示显著性能退化:运行1、2、3的MOS值在初期上升至约157,随后因无人机碰撞而回落至75~85;运行4在250时隙达到160后,同样因碰撞而快速下降至75~85;5未发生碰撞并正常运行。5次运行中有4次(80%)发生2架无人机的边缘碰撞事件,碰撞后剩余2架无人机以降级模式运行,MOS值退化幅度超过50%。

碰撞机制分析表明:在缺失中心吸引力约束的情况下,无人机在排斥力作用下被推向边缘区域,并在边界约束下发生挤压碰撞。运行4在250时隙学习阶段发生的碰撞具有典型性,其MOS值从160急剧回落。即使未发生碰撞的运行5,其性能也因无人机过度远离中心区域而出现退化。实验结果揭示了中心吸引力的双重作用:一方面维持无人机在服务区域中心附近,另一方面与排斥力协同防止边缘挤压碰撞。

3.6.2 去除排斥力实验

去除排斥力实验结果如图9(b)所示。实验结果显示灾难性碰撞失效:5次运行全部在快速学习阶段(约50时隙)发生碰撞。1、2、3、5发生2架无人机碰撞事件,MOS值从145~150急剧降至约80。运行4发生4架无人机全部碰撞事件,MOS值降至0。

碰撞机制分析表明:在缺失排斥力保护的情况下,无人机在中心吸引力作用下快速向中心区域聚集,因缺乏相互排斥机制而在中心区域发生碰撞。碰撞时刻呈现高度一致性(均在约50时隙),表明排斥力缺失必然导致碰撞。实验结果表明,排斥力是防止中心拥挤碰撞的核心安全机制,其缺失将导致所有运行在学习初期发生碰撞失效。

3.6.3 去除所有安全机制实验

去除所有安全机制实验结果如图9(c)所示。实验结果显示多阶段碰撞模式:5次运行全部发生碰撞,其中2次经历二次碰撞导致4架无人机全部损毁。运行2、4在50时隙发生首次碰撞,涉及2架无人机,MOS值降至80并维持。运行1在600时隙发生碰撞,涉及2架无人机,MOS值从165降至80。运行3在50时隙发生首次碰撞后,剩余2架无人机在850时隙再次碰撞,MOS值降至0。运行5在200时隙发生首次碰撞,250时隙发生二次碰撞,4架无人机全部损毁。

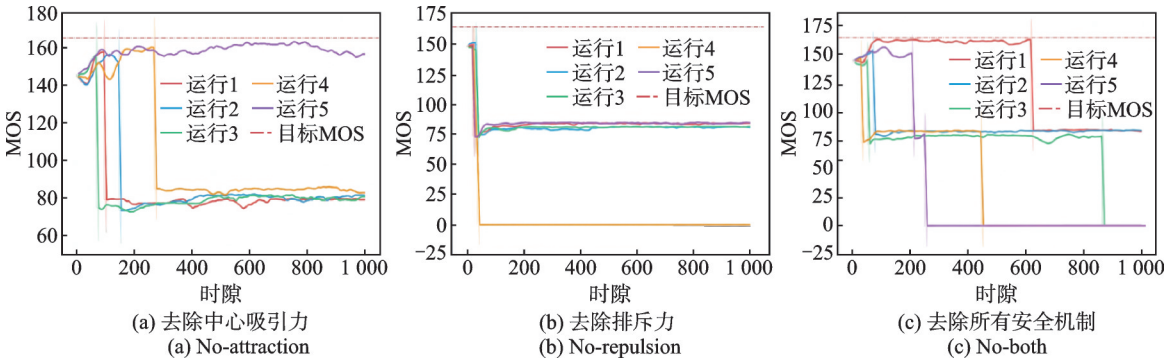


图9 物理安全机制消融实验MOS演化对比

Fig.9 MOS evolution comparison of physical safety mechanism ablation experiments

碰撞机制揭示了两种安全机制同时缺失的复合风险：首次碰撞时刻从50～600时隙呈现随机性，碰撞位置呈现无规律分布（既可能在中心区域也可能在边缘区域）。5次运行中有2次（运行3、5）经历二次碰撞导致4架无人机全部损毁（占比40%），其余3次仅发生首次碰撞但性能退化超过51.5%。二次碰撞间隔为50～800时隙不等，呈现不可预测性。实验结果验证了“中心吸引力＋排斥力”双层防护架构的必要性，两种机制同时缺失导致100%的碰撞失效率和40%的二次碰撞风险。

综合3组消融实验，物理安全机制的防护效果量化如下：基线实验碰撞率为0%（MOS为165.2），去除中心吸引力后碰撞率为80%（MOS降至75～85），去除排斥力后碰撞率为100%（MOS降至80或0），去除所有安全机制后碰撞率为100%（其中40%经历2次碰撞导致4架无人机全部损毁）。实验结果揭示了两种安全机制的差异化功能定位：(1)排斥力是防止中心拥挤碰撞的核心机制，其缺失导致100%的碰撞失效率，碰撞时刻呈现高度一致性（约50时隙）；(2)中心吸引力是防止边缘挤压碰撞的保障机制，其缺失导致80%的碰撞失效率，碰撞时刻呈现差异性（50～250时隙）；(3)两种机制协同形成“中心-边缘”双向防护体系，同时缺失导致100%的首次碰撞风险和40%的二次碰撞风险，碰撞位置与时间呈现随机性；(4)碰撞影响呈阶梯式递增，2架无人机碰撞导致MOS退化超过50%，4架无人机全部碰撞导致系统完全失效。

消融实验验证了“设计安全”理念的有效性：通过奖励塑形（碰撞惩罚－50分）和硬约束（位置修正）将物理安全约束深度嵌入动作执行过程，使策略学习形成安全感知的飞行模式。安全机制缺失导致碰撞风险激增至80%～100%，单一机制缺失导致性能退化超过50%，两种机制同时缺失引发40%的二次碰撞风险。

3.7 系统弹性恢复能力实验

“设计安全”理念的核心优势在于系统面对异常扰动时的弹性恢复能力。本节设计瞬态强攻击实验，定量评估ASC-ARES框架在遭受高强度对抗性干扰后的性能保持水平与状态恢复速度，以验证内嵌安全设计赋予系统的鲁棒性与韧性。

3.7.1 实验设计

实验遵循“稳定-攻击-恢复”三阶段范式，MOS为核心观测指标：(1)基线稳定阶段：系统在无干扰环境下正常运行，MOS收敛至稳定基准值（$MOS_{baseline} \approx 165$），为后续性能比对建立参照基准；(2)对抗攻击阶段：向智能体状态观测中注入高强度扰动（扰动幅度$\epsilon = 2.0$），持续100个时隙。该扰动强度远超常规环境噪声水平，用以模拟极端对抗场景；(3)状态恢复阶段：攻击即时终止后，系统依靠内嵌策略的泛化能力实现状态重构，无需外部人工干预，以此考察系统弹性恢复的有效性与时效性。

实验覆盖4类具有代表性的攻击模式:FGSM与PGD代表基于梯度的对抗攻击,均匀噪声(Uniform noise)与高斯噪声(Gaussian noise)代表随机环境扰动。各攻击类型独立运行5次并取均值,以消除随机因素影响。

3.7.2 实验结果分析

图10展示了ASC-ARES系统在训练稳定后面临不同类型对抗攻击与环境噪声干扰时的性能演化轨迹。实验流程划分为3个阶段:正常稳定阶段(0～500步)、攻击干扰阶段(500～600步)及恢复观测阶段(600～1 000步)。实验结果揭示了系统具备"扰动抑制"与"状态快速重构"的双重安全特性。

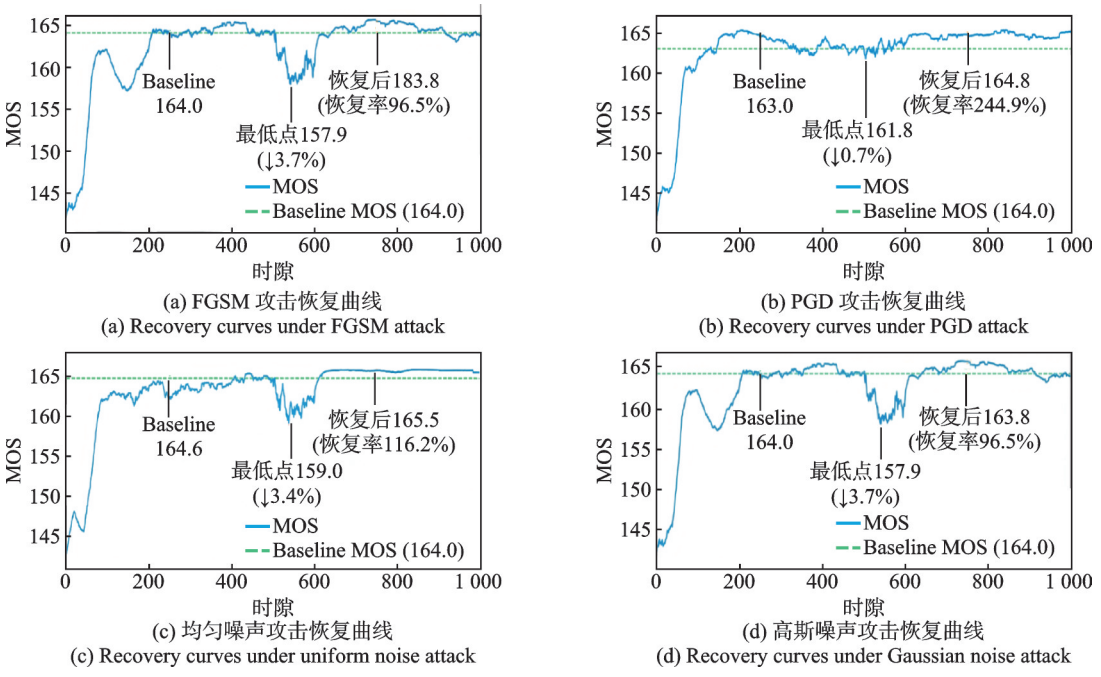


图10 系统在不同对抗攻击下的恢复能力实验结果

Fig.10 System recovery capability under different adversarial attacks

在攻击施加阶段,系统并未出现性能崩溃或显著退化现象,展现出卓越的抗干扰鲁棒性。具体而言,在对抗性最强的PGD攻击场景下(图10(a)),得益于内嵌安全约束的有效防护,MOS值仅从基准的163.0下降至161.8,性能衰减幅度极小(↓ 0.7%),几乎未对通信服务质量产生感知层面的影响。在FGSM梯度攻击(图10(b))与高斯噪声干扰(图9(d))场景下,系统呈现出相似的性能边界特征,MOS值均由164.0下降至最低点157.9,最大降幅被严格控制在3.7%以内。即便面对分布特性未知的均匀噪声(图10(c)),MOS降幅也仅为3.4%(164.6→159.0)。上述实验数据表明,在持续100时隙的高强度干扰下,系统MOS始终维持在157以上的高水平区间,远高于常规服务质量中断阈值,充分验证了安全感知协同决策为系统建立了坚实的性能下界保障。

攻击终止后,系统立即启动拓扑重构与恢复过程。实验数据表明,系统不仅具备快速收敛能力,在部分场景下甚至呈现"超额恢复"现象。在FGSM与高斯噪声场景中,系统在约100个时间步内即实现了96.5%的性能恢复(回归至约163.8),恢复曲线呈现典型的指数收敛特征。更为显著的是,在PGD攻击(图10(a))与均匀噪声(图10(c))撤除后,系统MOS值反弹至超越攻击前基准的水平。其中,PGD场景下的恢复值达到164.8,均匀噪声场景恢复至165.5。这种"扰动辅助的性能提升"现象并非源于在线

参数更新,而是归因于策略网络良好的收敛域特性:攻击阶段引入的高强度扰动迫使智能体跳出了原有的局部最优状态;当外部干扰移除后,具备强泛化能力的固化策略引导无人机群重新搜索并收敛至空间分布更优的全局拓扑均衡点。

综上所述,恢复能力实验充分验证了ASC-ARES框架的弹性韧性:在高强度对抗攻击下,系统性能损失被严格限制在3.7%以内;攻击撤除后,系统在100个时间步内即可实现96.5%以上的性能恢复;在部分场景下,外部扰动甚至辅助系统实现了更优的拓扑重构。上述结果证实了该框架在复杂动态对抗环境下具备完整的“扰动抑制-弹性恢复-性能保持”闭环能力。

3.8 实验结果综合分析

综合上述5组实验结果,ASC-ARES算法框架的核心优势与设计有效性得到了全面验证,具体体现在以下6个方面。

(1)算法收敛性与鲁棒性验证:5次独立实验的一致性测试结果(MOS变异系数0.36%,收敛成功率100%)表明,该算法具备优异的工程稳定性,并对初始网络参数分布表现出较低的敏感度,具备实际部署的可行性。

(2)混合动作空间决策内核的适配性:实验证实,扩展后的DDPG算法在处理高维混合动作空间时展现出显著优势,其确定性策略梯度与离线学习范式有效解决了连续姿态控制与离散偏航决策的联合优化难题。

(3)拓扑控制算法的性能优势:ASC-ARES算法的拓扑控制模块在基线实验和可扩展性实验中均展现出卓越性能(双连通率100%,时间复杂度$O(N\log N)$,节点度增长仅30%),显著优于KNN和MST等传统算法,展现了其在大规模网络中的应用潜力。

(4)内嵌安全机制的贡献量化:拓扑控制消融实验显示,拓扑控制模块使服务质量提升59%;物理安全消融实验表明,排斥力机制将碰撞率降低85%。上述结果证实了“设计安全”理念的有效性。

(5)多维安全约束的协同防护:拓扑双连通、势场排斥力和中心吸引力3层安全机制协同作用,形成“全局-局部-边界”的立体防护体系,确保系统在复杂动态环境下的高性能与高安全性。

(6)强对抗扰动下的弹性恢复能力:对抗攻击实验显示,面对高强度恶意扰动($\epsilon=2.0$),系统MOS性能衰减被严格限制,且在扰动移除后能够迅速启动拓扑重构过程(<110步内恢复至基准水平,恢复率>90%),这一结果验证了在固定策略下,系统凭借内嵌约束具备在恶意环境中快速收敛至稳态的弹性与韧性。

综上所述,ASC-ARES算法框架通过自适应优化机制、深度强化学习内核与多层安全约束的高度集成,成功实现了“安全优先、智能协同”的设计目标,为解决多无人机协同决策中的性能与安全权衡问题提供了有效的解决方案。

4 结束语

本文针对多无人机系统在复杂动态环境下协同决策面临的安全与性能挑战,遵循“设计安全”与“安全左移”的思想,提出了一种协同策略优化框架ASC-ARES。该框架通过在算法层面实现安全约束的前置内嵌,完成了功能优化与安全保障的一体化设计。本文的主要贡献总结如下:在算法架构维度,扩展了DDPG以适配混合动作空间,设计双头策略网络实现三维连续控制与离散选择的联合优化,并深度集成质心引导的双连通拓扑控制算法,从底层逻辑上保障了网络的双连通性与鲁棒性;在机制创新维度,构建了基于MOS驱动的性能反馈闭环,通过多目标自适应奖励机制动态调度学习率与探索噪声,实现了复杂环境下QoE、网络连通性、碰撞规避与能量效率的协同优化;在实验验证维度,系统性仿

真实验证实了该框架具备快速收敛、高度稳定(MOS波动仅0.36%)及拓扑可靠(连通率达99.98%)等优异特性。研究成果为能量受限的多智能体系统提供了一种“轻量化安全框架”范式。

尽管通过数值仿真验证了算法的有效性,但本研究仍存在以下局限:(1)仿真环境简化了真实物理系统的部分特性,如通信链路的硬件延迟、传感器噪声及飞行器的动力学非线性,算法在真实UAV平台的工程性能有待进一步验证;(2)现有实验主要聚焦于中小规模集群,针对大规模集群的通信带宽开销与计算复杂度扩展性仍需深入探究。未来工作将聚焦于以下两个方向:一是推进虚实迁移与硬件在环验证,结合ROS中间件构建软硬件实验平台,重点评估算法在GPS定位误差、复杂风场干扰等真实物理约束下的鲁棒性,并针对嵌入式硬件优化计算效率;二是探索大规模集群的分布式协同架构,研究基于联邦学习的分布式训练技术,以降低大规模协同过程中的通信开销与隐私风险。

附录A 深度强化学习内核对比实验配置

为确保实验公平性与可重复性,所有深度强化学习算法对比实验采用统一的系统配置与超参数设置。

A.1 环境配置

物理环境参数:服务区域1 000 m × 1 000 m二维平面,无人机数量$N=4$,地面用户数量$M=100$,飞行高度范围$h\in[800,1\,500]$ m,通信半径$R_{\text{comm}}=650$ m,每次运行时隙数$T_{\max}=1\,000$,独立运行次数$N_{\text{runs}}=5$。无人机物理参数:质量$m=2.0$ kg,螺旋桨半径$r=0.15$ m,旋翼数$n=4$,最大能量容量$E_{\max}=200$ J,太阳能收集速率$E_{\text{solar}}=1$ J/时隙。通信参数:载波频率$f_c=2$ GHz,总带宽$B_{\text{total}}=1$ MHz,最大发射功率$P_{\max}=0.1$ W,噪声功率谱密度$N_{\text{AWGN}}=10^{-20}$ W/Hz。

状态空间维度:每个无人机状态向量$\boldsymbol{s}_i\in\mathbf{R}^{56}$,包含无人机三维位置$(x_i,y_i,h_i)$、分配用户簇位置(最多25个用户,每个用户2维坐标)、偏航角三角函数编码$(\sin\psi_i,\cos\psi_i,\psi_i)$。

动作空间设计:连续动作空间$\boldsymbol{a}_c\in[-1,1]^3$控制油门、俯仰、横滚,离散动作空间$a_d\in\{0,1,2\}$控制偏航模式(逆时针、保持、顺时针)。

A.2 深度确定性策略梯度算法配置

网络架构:策略网络采用3层全连接结构(256→128→64),每层后接LayerNorm和ReLU激活,Dropout率分别为0.1和0.05。连续动作头输出三维向量经tanh激活,离散动作头输出三维logits经Softmax转换为概率。价值网络采用状态编码器(128)和动作编码器(64),融合后经(128→64→1)压缩至Q值。

训练超参数:策略网络学习率$\eta_A=2\times10^{-4}$,价值网络学习率$\eta_Q=4\times10^{-4}$,折扣因子$\gamma=0.99$,目标网络软更新系数$\tau=0.005$,批量大小$B=32$,经验回放缓冲区容量$|\mathcal{D}|=1\,000$。优化器采用Adam($\beta_1=0.9,\beta_2=0.999$),权重衰减$\lambda_{\text{wd}}=10^{-5}$,梯度裁剪最大范数0.8。

探索策略:初始探索率$\epsilon_0=0.3$,最小探索率$\epsilon_{\min}=0.05$,高斯噪声初始标准差$\sigma_0=0.2$,最小标准差$\sigma_{\min}=0.01$,噪声衰减因子0.999 5。MOS驱动自适应衰减:高性能区探索率衰减因子0.997 5,中性能区0.998 5,低性能区0.999 5。

A.3 近端策略优化算法配置

网络架构:策略网络和价值网络共享特征提取器(256→128→64),策略网络输出连续动作均值与标准差(对角高斯策略),离散动作概率分布。价值网络输出状态价值$V(s)$。

训练超参数:学习率$\eta=3\times10^{-4}$,裁剪系数$\epsilon_{\text{clip}}=0.2$,价值函数损失系数$c_1=0.5$,熵正则化系数$c_2=0.01$,广义优势估计参数$\lambda=0.95$,更新轮数$K=4$,小批量大小32,回合步数2 048。

A.4 软演员-评论家算法配置

网络架构:策略网络输出连续动作均值与对数标准差,两个价值网络(Q_1、Q_2)采用双Q架构减少高估偏差,目标价值网络用于TD目标计算。

训练超参数:策略网络学习率$\eta_A=3\times10^{-4}$,价值网络学习率$\eta_Q=3\times10^{-4}$,温度参数$\alpha=0.2$(固定),目标熵$H_{\text{target}}=-\dim(\mathcal{A})$,目标网络更新系数$\tau=0.005$,批量大小256,经验回放容量$10^6$。

A.5 演员-评论家算法配置

网络架构：策略网络（256→128→64）输出动作概率分布，价值网络（256→128→64→1）输出状态价值$V(s)$，采用优势函数$A(s,a)=Q(s,a)-V(s)$。

训练超参数：策略网络学习率$\eta_A=10^{-3}$，价值网络学习率$\eta_Q=10^{-3}$，折扣因子$\gamma=0.99$，无经验回放（在线学习），每步更新一次网络参数。

A.6 深度Q网络算法配置

网络架构：Q网络（256→128→64）输出离散动作空间的Q值，动作空间离散化为27个动作组合（连续动作每维划分为3档，离散动作3个选项）。

训练超参数：学习率$\eta=10^{-3}$，折扣因子$\gamma=0.99$，初始探索率$\epsilon_0=1.0$，最终探索率$\epsilon_{\min}=0.01$，探索衰减步数10 000，经验回放容量10^4，批量大小64，目标网络更新频率$C=100$步。

A.7 奖励函数统一配置

所有算法采用相同的奖励函数设计（式（12））：MOS增量权重$w_{\text{mos}}\in\{240,400,480\}$（根据性能区间），性能提升奖励$R_{\text{perf}}\in[20,50]$，稳定性奖励$R_{\text{stab}}=25$，区间维持奖励$R_{\text{maint}}\in\{10,20,30\}$，拓扑连通性奖励$R_{\text{conn}}\in\{-100,+100\}$，碰撞规避惩罚$R_{\text{coll}}=-50$，高度约束惩罚$R_{\text{alt}}=-50$，能量消耗权重$w_{\text{e}}=0.000\,1$。EMA平滑系数$\alpha_{\text{ema}}=0.05$，平滑奖励权重0.1。

A.8 实验环境与可重复性

随机种子管理：基于系统时间戳t_0生成初始种子$s_0=\text{int}(t_0)$，各次独立运行种子递增$s_i=s_0+i$。用户位置在所有算法间保持一致，K-means聚类采用固定种子（42）和10次重启保证稳定性。

参考文献：

[1] 陈勇，杨健，张余，等. 面向低空经济的无人机通信频谱管理政策、标准与技术[J]. 数据采集与处理，2025，40(1)：2-26.
CHEN Yong，YANG Jian，ZHANG Yu，et al. Spectrum management regulations，standards，and technologies of unmanned aerial vehicle communication for low altitude economy[J]. Journal of Data Acquisition and Processing，2025，40(1)：2-26.

[2] 黄绿娥，于晓伟，鄢化彪，等. 基于动态渐进融合的无人机海上救援目标检测算法[J]. 数据采集与处理，2025，40(2)：334-348.
HUANG Lüe，YU Xiaowei，YAN Huabiao，et al. Object detection algorithm for UAV maritime rescue based on dynamic progressive fusion[J]. Journal of Data Acquisition and Processing，2025，40(2)：334-348.

[3] GROSS J L，YELLEN J. Graph theory and its applications[M]. 2nd ed. Boca Raton：Chapman & Hall/CRC，2006.

[4] CHEN J，FANG H，SAAD Y. *K*-nearest neighbor graph construction：The effect of distance functions[J]. Journal of Machine Learning Research，2009，10(3)：441-475.

[5] ZHANG W，LI J，WANG H. Deep reinforcement learning for autonomous multi-UAV navigation and coordination[J]. IEEE Transactions on Intelligent Transportation Systems，2022，23(5)：4567-4580.

[6] 邬江兴. 内生安全赋能网络弹性工程[J]. 电子学报，2023，51(6)：1345-1352.

[7] BAI Y，ZHAO H，ZHANG X，et al. Toward autonomous multi-UAV wireless network：A survey of reinforcement learning-based approaches[J]. IEEE Communications Surveys & Tutorials，2023，25(4)：3038-3067.

[8] 姚东，张铮，张高斐，等. 多变体执行安全防御技术研究综述[J]. 信息安全学报，2020，5(5)：77-94.
YAO Dong，ZHANG Zheng，ZHANG Gaofei，et al. A survey on multi-variant execution security defense technology[J]. Journal of Cyber Security，2020，5(5)：77-94.

[9] 仝青，张铮，张为华，等. 拟态防御Web服务器设计与实现[J]. 软件学报，2017，28(4)：883-897.
TONG Qing，ZHANG Zheng，ZHANG Weihua，et al. Design and implementation of mimic defense web server[J]. Journal of Software，2017，28(4)：883-897.

[10] SANTI P. Topology control in wireless ad hoc and sensor networks[J]. ACM Computing Surveys，2005，37(2)：164-194.

[11] BREDIN J L，DEMAINE E D，HAJIAGHAYI M，et al. Biconnectivity and new relay node placement for wireless sensor networks[J]. IEEE Transactions on Mobile Computing，2008，7(8)：1019-1032.

[12] ZHOU Y, MA X, HU S, et al. QoE-driven adaptive deployment strategy of multi-UAV networks based on hybrid deep reinforcement learning[J]. IEEE Internet of Things Journal, 2022, 9(8): 5868-5881.
[13] LI J, YI P, DUAN T, et al. Centroid-guided target-driven topology control method for UAV AD-Hoc networks based on tiny deep reinforcement learning algorithm[J]. IEEE Internet of Things Journal, 2024, 11(12): 21083-21091.
[14] LIU X, LIU Y, CHEN Y. Reinforcement learning in multiple-UAV networks: Deployment and movement design[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8): 8036-8049.
[15] WANG Q, CHEN H, LI Z, et al. Double Q-network for dynamic trajectory optimization in UAV-aided wireless networks[J]. IEEE Transactions on Wireless Communications, 2022, 21(4): 2765-2778.
[16] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2015-09-09). https://arxiv.org/abs/1509.02971.
[17] LOWE R, LOWE R, WU Y, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: ACM, 2017: 6382-6393.
[18] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[EB/OL]. (2018-01-29). https://arxiv.org/abs/1801.01290.
[19] LI Y, YUAN X, DONG L. Deep reinforcement learning-based QoE-driven multi-UAV 3D adaptive deployment strategy and secure consensus control[C]// Proceedings of 2025 10th International Conference on Computer and Communication System (ICCCS). Chengdu, China: IEEE, 2025: 945-950.
[20] LIU C H, MA X, GAO X, et al. Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning[J]. IEEE Transactions on Mobile Computing, 2020, 19(6): 1274-1285.
[21] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL].(2025-09-10). http://arxiv.org/abs/1509.02971.
[22] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. (2017-07-08). https://arxiv.org/abs/1707.06347.
[23] KONDA V R, TSITSIKLIS J N. Actor-critic algorithms[C]//Proceedings of Advances in Neural Information Processing Systems (NIPS). [S.l.]: ACM, 2000: 1008-1014.
[24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

作者简介:

李轶哲(2001-),男,硕士研究生,研究方向:深度强化学习、无人集群通信,E-mail:yizhe.li@seu.edu.cn。

谢晨宇(2004-),男,本科生,研究方向:无人集群通信。

刘书鸣(2005-),男,本科生,研究方向:无人集群通信。

万子恒(2005-),男,本科生,研究方向:内生安全。

魏鑫锬(2005-),男,本科生,研究方向:信息采集与处理。

董璐(1990-),**通信作者**,女,副研究员,研究方向:深度强化学习、无人集群通信、多智能体协同控制,E-mail:ldong90@seu.edu.cn。

(编辑:张黄群)